LA-UR-20-20207
Approved for public release; distribution is unlimited.

Title: Visual Analytics for Large Scale Scientific Simulations, Fiscal Year 2019

Author(s): Woodring, Jonathan Lee
Shen, Han Wei
Peterka, Tom

Intended for: Report

Issued: 2020-01-13 (rev.2)

Visual Analytics for Large Scale Scientific Ensemble Datasets

Fiscal Year 2019, LA-UR-20-20207

Jonathan Woodring, Han-Wei Shen, and Tom Peterka

Introduction

Scientific ensemble data sets have played increasingly more important roles for uncertainty quantification in various scientific and engineering domains, such as climate, weather, aerodynamics, and computational fluid dynamics. Ensembles are collections of data produced by simulations or experiments conducted with different initial conditions, parameterizations, or phenomenological models. They are usually used to describe complex systems, study sensitivities to initial conditions and parameters, and mitigate uncertainty. The goal of this proposal is to develop visual analytic techniques for large scale scientific ensemble data sets. Using ensemble simulations as an example, for a single run of such a simulation, there can be data generated in the range of several hundred gigabytes to tens of terabytes. A large scale ensemble dataset can consist of hundreds or thousands of such instances, with many variables in the form of scalar, vector, or tensor, and has a large number of samples in the high-dimensional input parameter space.

We proposed to research and develop methods for large-scale data analytics and visualization as applied to scientific data ensembles in several different topic areas: 1) *Exploration of Local Uncertainty with Distributions*, 2) *Exploration and Tracking of Ensemble Features*, and 3) *Exploration of Multivariate Ensemble Parameters*. Additionally, we proposed to tackle the scalability of these methods as applied towards DOE applications of interest: 1) *Automation of In Situ Ensemble Analytics* and 2) *Domain Specific and Laboratory Applications*. Below, we present selected results of our efforts in each of these aforementioned areas for FY 2019.

Exploration of Local Uncertainty with Distributions

NNVA: Neural Network Assisted Visual Analysis of Yeast Cell Polarization Simulation [1]

Best Paper, Honorable Mention – IEEE VAST 2019

Computational models are designed to simulate real-world physical phenomena in many scientific disciplines. However, these models tend to be computationally

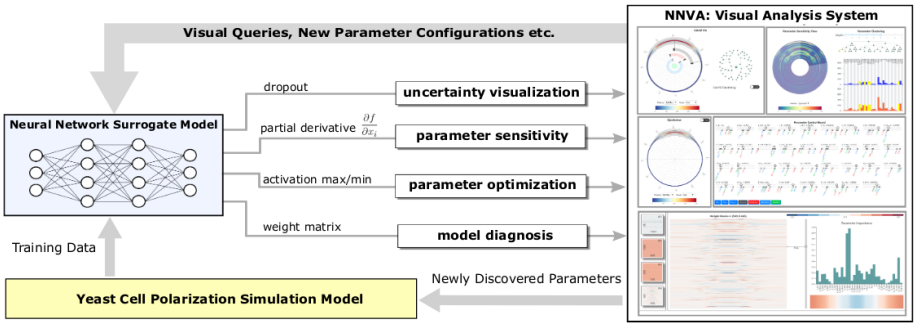


Figure 1: *A neural network-based surrogate model acts as the analysis framework, driving our interactive visual analysis system for analyzing a computationally expensive yeast simulation model.*

very expensive and involve a large number of simulation input parameters, which need to be properly calibrated before the models can be used for studies.

We proposed a visual analysis system to facilitate interactive exploratory analysis of the high-dimensional input parameter space for a yeast cell polarization simulation. The system can assist computational biologists to visually calibrate the input parameters by modifying the parameter values. They are able to immediately visualize the predicted simulation outcome without needing to run the simulation for every instance.

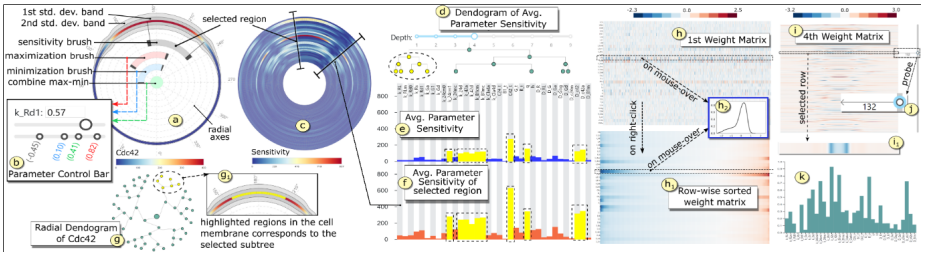


Figure 2: *Primary Visualizations and Interaction techniques: (a) Predicted Cdc42 concentration across the membrane along with uncertainty bands and selection brushes. (b) Parameter control bar. (c) Spatial parameter sensitivity. (d) Linear cluster tree for average parameter sensitivity. (e,f) Average parameter sensitivities. (g) Radial cluster tree for predicted Cdc42. (h) First weight matrix. (i) Final weight matrix. (j) Row selectionprobe. (k) Average parameter sensitivity for selected pattern.*

Our visual analysis system is driven by a trained neural network-based surrogate model as the analysis framework. We demonstrated the advantage of using surrogate models for visual analysis by incorporating some of the recent advances in the field of uncertainty quantification, interpretability and explainability of

neural network-based models. The trained network is able to perform interactive parameter sensitivity analysis of the simulation, as well as, recommend optimal parameter configurations. We also facilitate detailed analysis of the trained network to extract useful insights about the model.

Exploration and Tracking of Ensemble Features

DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation [6]

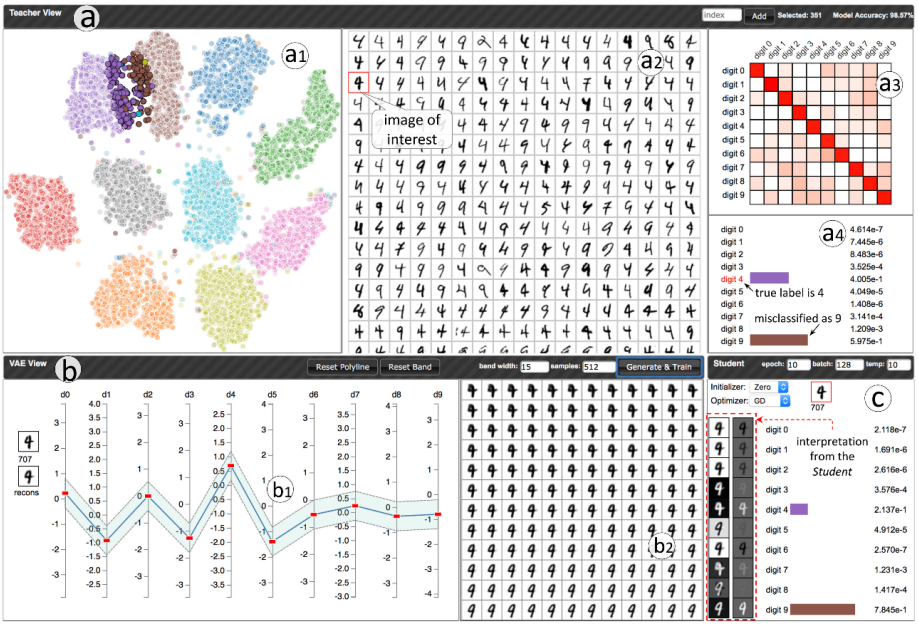


Figure 3: *DeepVID. (a) Teacher view: explore test data and assess the Teacher's performance with the t-SNE view (a1); Image-Grid view (a2); Confusion Matrix view (a3); Probability Distribution view (a4). (b) VAE view: explore the latent space with a parallel coordinates plot (b1); and the generated neighbors with the Neighbors view (b2). (c) Student view: visualize the trained Student for interpretation*

Deep Neural Networks (DNNs) have been extensively used in multiple disciplines due to their modeling performance. However, DNNs are considered black-boxes and the interpretation of their internal working and modeling mechanisms are challenging. Given that trust of a model is built on the understanding and intuition on how the model works, building tools for the interpretation of DNNs is of utmost importance; especially in their expanding use in safety-critical

applications (e.g., medical diagnosis, autonomous driving).

We proposed DeepVID, a Deep learning approach to Visually Interpret and Diagnose DNN models. In particular, we studied approach for image classifiers. We train a small locally-faithful surrogate models to mimic the behavior of a larger DNN around a particular datum of interest. These local models are generated to be sufficiently simple such that it can be visually interpreted (e.g., a linear model) in our system.

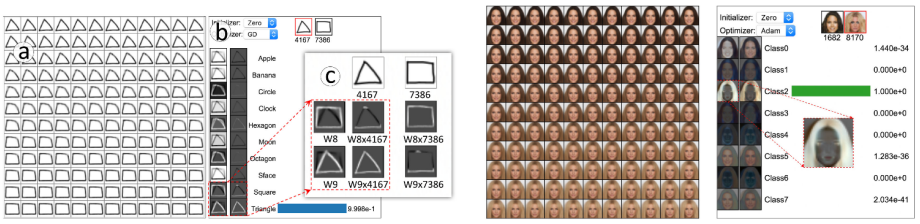


Figure 4: *Left: How a triangle and a square are differentiated. Right: DeepVID identifies the blond hair feature in image 8170.*

Knowledge distillation is used to transfer the model knowledge encoding from the larger DNN into the smaller surrogate model. Then, a deep generative model (i.e., variational auto-encoder) is used to generate data neighbors around a particular datum of interest. The neighbors, which contain small feature variances and semantic meanings, probe and explore the DNN's modeling behaviors around the datum of interest. Through comprehensive evaluations, as well as case studies conducted together with deep learning experts, we tested the effectiveness of our system.

Exploration of Multivariate Ensemble Parameters

InSituNet: Deep Image Synthesis for Parameter Space Exploration of Ensemble Simulations [2]

Best Paper – IEEE SciVIS 2019

We proposed InSituNet, a deep learning based surrogate model to support parameter space exploration for ensemble simulations that are visualized by in situ processing. In situ visualization, i.e., generating visualizations at simulation time, is used for the analysis of large-scale simulation data due to I/O and storage limitations. However, typical in situ visualization lacks the flexibility of post-hoc data exploration since raw data are no longer available. Multiple in situ image-based visualization approaches have been proposed to mitigate

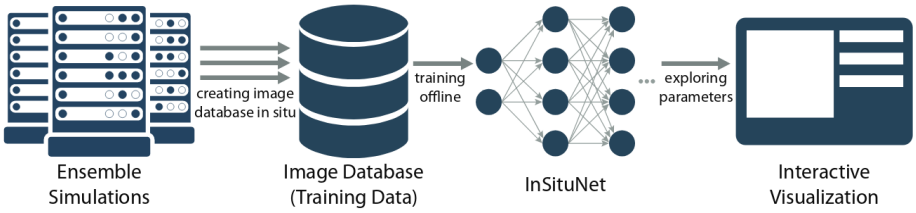


Figure 5: *The workflow of InSituNet. Simulations are run with various different simulation parameters on supercomputers, and visualization images are generated in situ for the different visual and view parameters. These generated images are collected into an image database. Then, a deep image synthesis model (i.e., InSituNet) is then trained, based on the collected data, which is used for parameter space visual exploration through an interactive interface.*

the flexibility limitation, but those approaches lack the ability to explore the simulation parameter space.

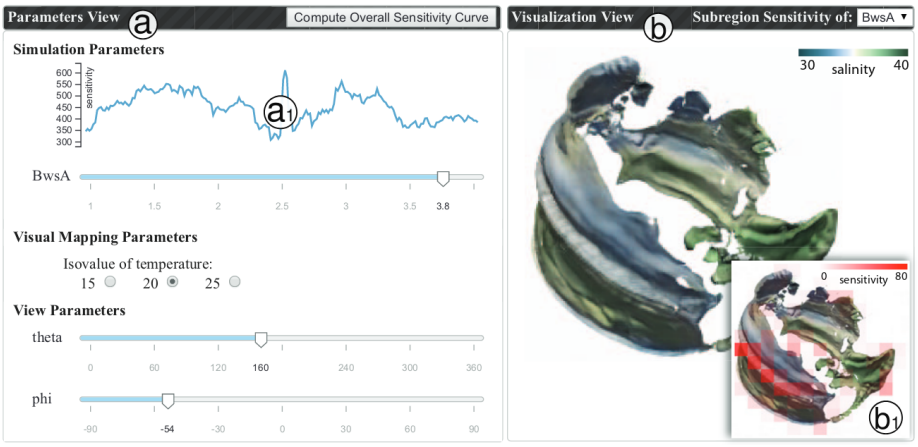


Figure 6: *Our visual interface for parameter space exploration. (a) The three groups of parameters: simulation, visual mapping, and view parameters. (b) The predicted visualization image and the sensitivity analysis result.*

Our approach provides the exploration of the parameter space by taking advantage of the recent advances in deep learning. Specifically, we designed InSituNet as a convolutional regression model to learn the mappings from simulation and visualization parameters to the visualization results. With our trained InSituNet model, analysts are able to generate many proxy images for any simulation parameter and visualization settings, enabling the flexible, exploratory analysis of an ensemble parameter space. We demonstrated the effectiveness of our visualization model with combustion, cosmology, and ocean simulations.

Automation of In Situ Ensemble Analytics

Statistical Super Resolution for Data Analysis and Visualization of Large Scale Cosmological Simulations [8]

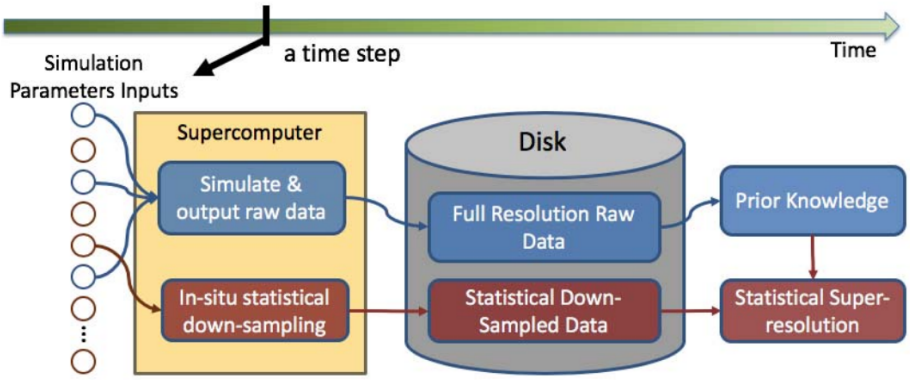
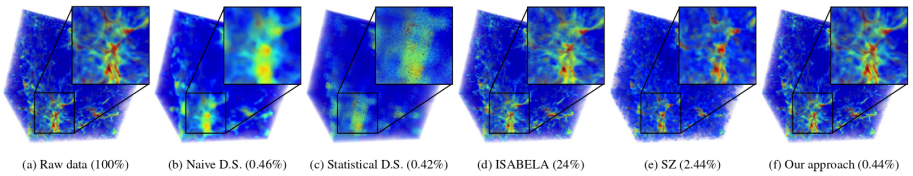


Figure 7: *An overview of Statistical Super Resolution.*

Cosmologists build models for studying the evolution of the universe, using different initial parameters for simulations. By exploring the outputs from different simulation runs, cosmologists approach an understanding of the evolution of our universe and its initial conditions. A modern, high-resolution cosmological simulation can generate dataset sizes on the order of petabytes. Thus, moving the datasets from the supercomputers to data analysis machines is infeasible to support exploratory visualization and analysis.

We proposed a novel approach called statistical super-resolution that tackles the big data problem for cosmological simulation analysis and visualization. It uses the datasets from a few high-resolution cosmology simulations to create a prior knowledge database. This database captures a relationship between smaller, low-resolution statistical information and larger, high-resolution data (simulation grid) representations of cosmological simulations.

At simulation run-time, we apply an in situ down-sampling to the create low-resolution statistical data representations from simulation runs, which minimizes the requirements of I/O bandwidth and storage. Then during visualization and analysis, high-resolution datasets are reconstructed from the low-resolution, statistical data, by using our prior knowledge database, to provide reliable scientific data analysis and high-quality visualizations.



(a) Raw data (100%) (b) Naive D.S. (0.46%) (c) Statistical D.S. (0.42%) (d) ISABELA (24%) (e) SZ (2.44%) (f) Our approach (0.44%)

Figure 8: *We compare the volume rendering of density quantity at time step 180 from a Nyx simulation run with input parameters: comovingh=0.61666, comovingOmB = 0.02216, and comovingOmM = 0.14328. D.S. stands for down-sampling. The numbers in sub-figure captions are the ratio of the representation storage consumption compared to the raw data size.*

Domain Specific and Laboratory Applications

Microparticle cloud imaging and tracking for data-driven plasma science [10]

Instrument (Image set)	**Application**	**Parameter** (feature)	**Data Model**
(simulations)[36]	fluids	position	*KNN* (*SOM*)
microscope[38]	cellular dynamics	position	*HF*
microscope[39]	self-assembly	cluster (distance)	*DM*
video[40]	surveillance	object	*CNN+RNN*
cryo-EM[41]	macromolecules	object	*deep CNN*
particle detector[42]	high-energy physics	track	*LSTM+CNN*
holograms[43]	colloidal science	3D position	*CC,CNN*
MNIST[44]	computer science	position cloud	*SO-Net*

Figure 9: *Examples of data-driven tracking algorithms and its applications. Illumination is assumed to be optical by default and otherwise specified. CC: Cascade classifier. CNN: convolutional neural network. DM: diffusion maps. HF: Haar features. KNN: Kohonen neural network. LSTM: Long Short Term Memory. RNN: recurrent neural network. SOM: self-organizing map.*

Large data sets give rise to a "fourth paradigm" of scientific discovery and technology development, extending other approaches based on human intuition, laws of physics, statistics, and computation. Both experimental and simulation data sets are growing in plasma science and technology, motivating the need for

applying data-driven discoveries.

We described our recent progress in microparticle cloud imaging and tracking (mCIT, uCIT) for laboratory plasma experiments. There are three types of microparticle clouds described: from exploding wires, in dusty plasmas, and in atmospheric plasmas. Experimental data sets were obtained with one or more imaging cameras, with up to 100k frames per second (fps). Analyses of the images generate time-dependent microparticle trajectories with time-dependent, two-dimensional or three-dimensional information about particle motion and ambient environment. These experiments generate massive image and particle track data, and motivated the development of machine-learning (ML) techniques for information extraction: a physics-constrained motion tracker, a Kohonen neural network (KNN) or self-organizing map (SOM), the feature tracking kit (FTK), and U-Net.

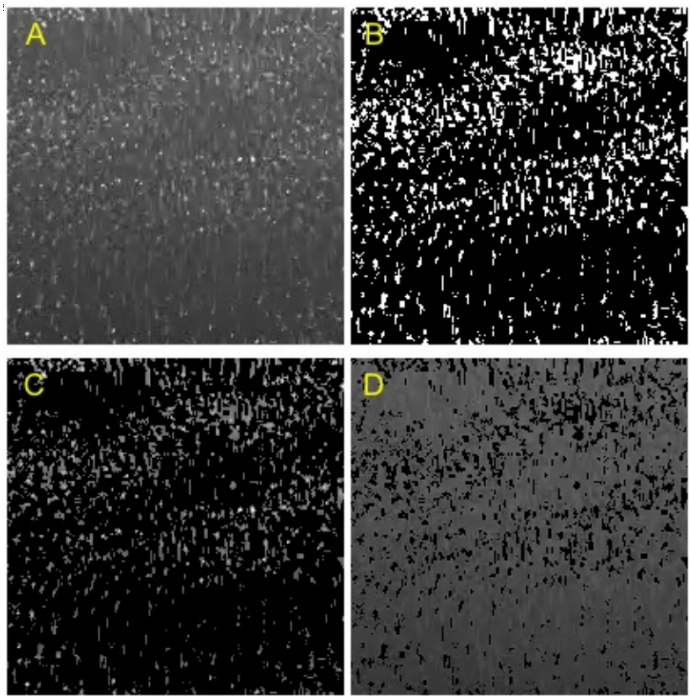


Figure 10: *Supervised noise reduction of the dusty plasma image set using U-Net. (A). An original image; (B). U-Net generated binary mask; (C). Masked original image; (D). Background after the subtraction of the masked image (C) from the original image (A).*

These methods were compared with each other particle tracking methods, where particle density and signal-to-noise ratio are as two important factors that affect the tracking accuracy. Fast Fourier transform (FFT) was used to reveal how

U-Net, a deep convolutional neural network (CNN), achieves the improvements for noisy scenes. The fitting parameters for a simple polynomial track model have been found to group into clusters, which reveal the geometric information about the camera setup. The mCIT or uCIT techniques, when enhanced with the data models, can be used to study the microparticle- or Debye-length scale plasma physics.

Citations

[1] Subhashis Hazarika, Haoyu Li, Ko-Chih Wang, Han-Wei Shen, and Ching-Shan Chou. *NNVA: Neural Network Assisted Visual Analysis of Yeast Cell Polarization Simulation.* IEEE Transactions on Visualization and Computer Graphics 26 (1), 34-44 (2020).

[2] Wenbin He, Junpeng Wang, Hanqi Guo, Ko-Chih Wang, Han-Wei Shen, Mukund Raj, Youssef S. G. Nashed, and Tom Peterka. *InSituNet: Deep Image Synthesis for Parameter Space Exploration of Ensemble Simulations.* IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE VIS 2019), 26(1):23-33, 2020.

[3] Xiaonan Ji, Han-Wei Shen, Raghu Machiraju, Alan Ritter, and Po-Yin Yen. *Visual Exploration of Neural Document Embedding in Information Retrieval: Semantics and Feature Selection.* IEEE Transactions on Visualization and Computer Graphics 25 (6), 2181-2192.

[4] Cheng Li and Han-Wei Shen. *Object-in-Hand Feature Displacement with Physically-Based Deformation.* Pacific Visualization Symposium (PacificVis), 2019 IEEE, 21-30.

[5] Xin Liang, Hanqi Guo, Sheng Di, Franck Cappello, Mukund Raj, Chunhui Liu, Kenji Ono, Zizhong Chen, and Tom Peterka. *Towards feature preserving 2D and 3D vector field compression.* In Proceedings of IEEE Pacific Visualization Symposium, 2020 (conditionally accepted).

[6] Junpeng Wang, Liang Gou, Wei Zhang, Hao Yang, and Han-Wei Shen. *DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation.* IEEE Transactions on Visualization and Computer Graphics 25 (6), 2168-2180.

[7] Ko-Chih Wang, Tzu-Hsuan Wei, Naeem Shareef, and Han-Wei Shen. *Ray-based Exploration of Large Time-varying Volume Data Using Per-ray Proxy Distributions.* IEEE Transactions on Visualization and Computer Graphics, 2019 (Early Access).

[8] Ko-Chih Wang, Jiayi Xu, Jonathan Woodring, and Han-Wei Shen. *Statistical Super Resolution for Data Analysis and Visualization of Large Scale Cosmological*

Simulations. Pacific Visualization Symposium (PacificVis), 2019 IEEE, 303-312.

[9] Yang Wang, Minzhu Yu, Guihua Shan, Han-Wei Shen, and Zhonghua Lu. *VISPubComPAS: a comparative analytical system for visualization publication data.* Journal of Visualization 22 (5), 941-953 (2019).

[10] Zhehui Wang, Jiayi Xu, Yao E. Kovach, Bradley T. Wolfe, Edward Thomas Jr., Hanqi Guo, John E. Foster, and Han-Wei Shen. *Microparticle cloud imaging and tracking for data-driven plasma science.* arXiv:1911.010000 [physics.plasm-ph], 2019.

Talks

- Soumya Dutta, Hanqi Guo, Hans-Christian Hege, and Han-Wei Shen. *Tutorial: Statistical data representation, visualization, and uncertainty analysis.* IEEE VIS 2019, October 21, 2019, Vancouver, BC, Canada.
- Han-Wei Shen. *An end-to-end in situ data analysis and visualization pipeline.* IEEE Pacific Visualization 2019 keynote.

Awards

- **Best Paper** (IEEE SciVIS 2019) – Wenbin He, Junpeng Wang, Hanqi Guo, Ko-Chih Wang, Han-Wei Shen, Mukund Raj, Youssef S. G. Nashed, and Tom Peterka. *InSituNet: Deep Image Synthesis for Parameter Space Exploration of Ensemble Simulations.* IEEE Transactions on Visualization and Computer Graphics (Proc. IEEE VIS 2019), 26(1):23-33, 2020.
- **Best Paper, Honorable Mention** (IEEE VAST 2019) – Subhashis Hazarika, Haoyu Li, Ko-Chih Wang, Han-Wei Shen, and Ching-Shan Chou. *NNVA: Neural Network Assisted Visual Analysis of Yeast Cell Polarization Simulation.* IEEE Transactions on Visualization and Computer Graphics 26 (1), 34-44 (2020).

Supported Students

Robert Gross is a Ph.D. student in Computer Science at the The Ohio State University, supervised by Prof. Han-Wei Shen. His research interests are in machine learning algorithms and visual analytics. He spent 10 weeks at the Los Alamos National Laboratory as a summer student working on data reduction and compression techniques for nuclear non-proliferation sensor data using machine learning and data modeling.

Subhashis Hazarika was a Ph.D. student supervised by Prof. Han-Wei Shen at the Ohio State University. He graduated in December 2019. His research interests are in distribution-based data modeling, ensemble data analysis, uncertainty visualization, and statistical methods.

Wenbin He was a Ph.D. student supervised by Prof. Han-Wei Shen at the Ohio State University. He graduated in December 2019. His research interests are in ensemble data visualization, machine learning, and statistical methods.

Xin Liang is a Ph.D. candidate in Computer Science at the University of California, Riverside. He received a B.S. degree in Computer Science from Peking University in 2014, with a minor in Math and Applied Math. His research primarily focuses on high performance computing, parallel and distributed systems, fault tolerance, data management and reduction, scientific visualization, and data analytics. He did two internships at Los Alamos National Laboratory and Pacific Northwest National Laboratory in 2017, working on container-based application encapsulation and scalable deep learning algorithms, respectively. His long-term internship in Argonne National Laboratory focuses on improving error-bounded lossy compressors under the guidance of Dr. Franck Cappello and Dr. Sheng Di, as well as designing feature-preserving lossy compressor with Dr. Thomas Peterka and Dr. Hanqi Guo in the last summer.

Jingyi Shen is a Ph.D. student in Computer Science at the The Ohio State University, supervised by Prof. Han-Wei Shen. Her research interests are in machine learning algorithms and visual analytics. She spent 10 weeks at the Los Alamos National Laboratory as a summer student working on data reduction and compression techniques for large-scale nuclear non-proliferation sensor data using machine learning and data modeling.

Jiayi Xu is a Ph.D. student supervised by Prof. Han-Wei Shen at the Ohio State University. His research interests are in data visualization and graph analysis. He spent 14 weeks at Argonne as a summer student working on scalable union-find algorithms for feature tracking.